DOCUMENT RESUME

ED 400 336 TM 025 762

AUTHOR Fan, Xitao; And Others
TITLE The Effects of Sample Size, Estimation Methods, and Model Specification on SEM Indices.
PUB DATE 9 Apr 96
NOTE 61p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Goodness of Fit; *Mathematical Models; Monte Carlo Methods; *Sample Size; *Structural Equation Models
IDENTIFIERS Covariance Matrices; *Specification Error

ABSTRACT

A Monte Carlo simulation study was conducted to investigate the effects of sample size, estimation method, and model specification on structural equation modeling (SEM) fit indices. Based on a balanced 3x2x5 design, a total of 6,000 samples were generated from a prespecified population covariance matrix, and eight popular SEM fit indices were studied. Two primary conclusions were suggested. First, for misspecified models, some fit indices appear to be noncomparable in terms of the information they provide about model fit; some fit indices also seem to be more sensitive to model misspecification. Second, estimation method strongly influenced almost all the fit indices examined, especially for misspecified models. These two issues do not appear to have been well documented in the previous literature. Perhaps the focus of most previous simulation studies on correctly specified models may have failed to detect these dynamics. It is further suggested that future research should study not only different models relative to model complexity, but also a wider range of model specification conditions, including correctly specified models as well as models specified incorrectly to varying degrees. (Contains 2 figures, 6 tables, and 26 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 336

fitstat.wp1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
BRUCE THOMPSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

The Effects of Sample Size, Estimation Methods, and Model Specification on SEM Fit Indices

Xitao Fan

Department of Psychology
Utah State University
Logan, Utah 84322-2810

Phone: (801)797-1451
Fax: (801)797-1448
E-Mail: FAFAN@CC.USU.EDU

Lin Wang

Bruce Thompson

Department of Educational Psychology
Texas A&M University
College Station, Texas 77843-4225

Running Head: SEM Fit Indices

Note: Please address correspondence about this article to: Xitao Fan, Ph.D., Department of Psychology, Utah State University, Logan, Utah 84322-2810.

Paper presented at the annual meeting of American Educational Research Association (session #16.52), April 9, 1996, New York City, NY.

TM025762

BEST COPY AVAILABLE

Abstract

A Monte Carlo simulation study was conducted to investigate the effects of sample size, estimation method, and model specification on SEM fit indices. Based on a balanced 3×2×5 design, a total of 6,000 samples were generated from a prespecified population covariance matrix, and eight popular SEM fit indices were studied. Two primary conclusions were suggested. First, for misspecified models, some fit indices appear to be non-comparable in terms of the information they provide about model fit; some fit indices also seem to be more sensitive to model misspecification. Second, estimation method strongly influenced almost all the fit indices examined, especially for misspecified models. These two issues do not appear to have been previously well documented in the literature. Perhaps the focus of most previous simulation studies on correctly specified models may have failed to detect these dynamics. It is further suggested that future research should not only study different models viz a viz model complexity, but also study a wider range of model specification conditions, including correctly specified models as well as models specified incorrectly to varying degrees.

Covariance structure analysis, or structural equation modeling (SEM), has been heralded as a unified model that joins methods from econometrics, psychometrics, sociometrics, and multivariate statistics (Bentler, 1994). The generality and wide applicability of structural equation modeling have been amply demonstrated (Bentler, 1992; Jöreskog & Sörbom, 1989). In recent years, SEM has become an increasingly popular statistical tool for researchers in psychology, education, and in the social and behavioral sciences in general. For researchers in these areas, SEM has become an important tool for testing theories with both experimental and non-experimental data (Bentler & Dudgeon, 1996). But despite its popularity in a variety of research situations, some thorny issues still haunt SEM applications. One such prominent issue is the assessment of model fit.

The assessment of model fit in SEM was initially framed within the dichotomous decision process of hypothesis testing: the model was either accepted as providing good fit to the data, or the model was rejected as fitting the empirical data poorly. The decision to accept or reject the hypothesis of fit was based on the probability level associated with the χ^2 value, which assesses the discrepancy between the original sample covariance matrix and the covariance matrix reproduced based on model specifications.

As is the case with statistical significance testing in general (Thompson, 1996), such an assessment of model fit is confounded with sample size: the power of the test increases with increases in the sample size used in the analysis. As a result,

model fit assessment becomes very stringent when sample size is large, and a minimal discrepancy between the original sample covariance matrix and the reproduced covariance matrix will be declared statistically significant, and consequently, rejected as having poor fit with the empirical data. But when sample size is small, the statistical test is lenient, and the test may fail to detect meaningful differences between the sample covariance matrix and the covariance matrix reproduced from the specified model.

Indices for Assessing Model Fit

Due to the generally recognized unsatisfactory nature of χ^2 statistic for model fit assessment (Thompson & Daniel, 1996), a variety of alternative indices for assessing model fit have been developed. Although some indices have been based on different theoretical rationales (Maiti & Mukherjee, 1991; Tanaka, 1993), many of them are superficially similar from a practical point of view. To get a sense about the number and variety of these indices, we only need to have a quick look at the output of current computer programs for SEM analysis. The SEM procedure under SAS (SAS Institute, 1990), PROC CALIS, outputs close to two dozen fit indices. Following the same trend, the new version of LISREL program (LISREL Mainframe Version 8.12) has substantially increased the number and type of fit indices in its output. Clear guidelines are currently lacking as regards the comparability and relative performance of these indices under different conditions. This somewhat chaotic state of affairs leaves many researchers confused about which indices to consult or present in their research work.

The main reason for this situation is that different types of fit indices were developed under different theoretical rationales, and there does not seem to exist one fit index which meets all our expectations for an ideal fit index, even if we had a complete consensus regarding our expectations. Although different opinions have been expressed as to what characteristics an ideal fit index should possess (Cudeck & Henly, 1991; Tanaka, 1993), an ideal fit index, as discussed by Gerbing and Anderson (1993), may possess at least three characteristics: (a) has a range between 0 and 1, with 0 indicating complete lack of fit, and 1 indicating perfect fit; (b) is independent of sample size; and (c) has known distributional properties to assist interpretation.

Since SEM fit indices were developed with different rationales, they may differ across several dimensions. Tanaka (1993) proposed a six-dimension typology for SEM fit indices, and attempted to categorize some popular fit indices along these six dimensions. This multifaceted nature of fit indices not only makes the comparison among fit indices difficult, but also makes it nearly impossible to select the "best" index from all those available.

Statistically, most popular fit indices fall into one of several types, and they were developed with different motivations (Gerbing & Anderson, 1993). The first type of fit index--*covariance matrix reproduction* indices--attempts to assess the degree to which the reproduced covariance matrix based on the specified model has accounted for the original sample covariance

matrix. This type of fit index can be conceptualized as the multivariate counterpart of the coefficient of determination (R^2) as in regression or ANOVA analysis (Tanaka & Huba, 1989). Examples of this type of fit indices are the Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit Index (AGFI) (Jöreskog & Sörbom, 1989).

The second type of fit index--*comparative model fit* indices--assess model fit by evaluating the comparative fit of a given model with that of a more restricted null model. In practice, the null model is usually a model which assumes no relationship among the indicators of the model. Reservations have been expressed about the appropriateness of using such null models as comparative baselines (Sobel & Bohrnstedt, 1985). Bentler and Bonnet's normed and non-normed fit indices (NFI and N_NFI) and Bollen's incremental fit index (DELTA2) both belong to this family.

The third type of fit index--*parsimony weighted* indices--specifically takes model parsimony into consideration by imposing penalties for specifying more elaborate models. More specifically, these fit indices consider both model fit and the degrees of freedom used for specifying the model. If good model fit is obtained at the expense of freeing more parameters, a penalty will be imposed. The reasoning in this type of model assessment is embedded in the long tradition of science going back to William of Occam's razor: between two models that fit data equally, the simpler model is more likely to be true, and therefore is also more likely to be replicated. Besides, statistically, better fit is

always obtained when more parameters in the model are freed. The parsimony indices proposed by James, Mulaik and Brett (1982) and by Mulaik et al. (1989) represent this type.

A recent development in model fit assessment makes use the noncentrality statistic from the noncentral χ^2 distribution for constructing fit indices. Based on sample noncentrality statistic, McDonald (1989) proposed an index of noncentrality. Bentler (1990) proposed the Comparative Fit Index (CFI) which also uses the sample noncentrality statistic. As with other fit indices proposed by Bentler, CFI assesses model fit relative to a baseline null model.

Some Considerations for Assessing Fit Indices

As discussed before, one major problem caused by the variety of SEM fit indices is that they create confusion in research practice. Not only are the rationales for different indices unclear to many researchers, but clear guidelines are also lacking as regards choosing among these indices. Furthermore, most fit indices have unknown distributional properties, thus making interpretation of sample fit indices very difficult.

The obvious reason for the lack of clear guidelines for choosing among different indices is that we simply do not fully understand the performance characteristics of these indices under different conditions. Due to the multifaceted nature of fit indices, and to different rationales for developing these indices, there does not seem to be a straightforward criterion against which the performance of all fit indices can be judged. Although it is not realistic to expect one straightforward criterion for judging

the performance of fit indices, several related criteria can be considered for this purpose.

First, despite the arguments in support of the role sample size plays in statistical decisions (Cudeck & Henly, 1991), the fact that the development of many fit indices was motivated to overcome the shortcomings of χ^2 statistic, especially its sensitivity to sample size, cannot be ignored. For this reason, ideally, fit indices should be insensitive to or independent of sample size (Bollen, 1986). This means that an index's variation contributed by *sample size* conditions should be as small as possible.

Second, under the assumption of multivariate normality, model fitting and estimating can be accomplished through different statistical procedures, such as maximum likelihood (ML) or generalized least squares (GLS). Ideally, fit indices should be invariant over this condition, i.e., different statistical theories should not result in excessively variable indices for the same data. This reasoning leads to the expectation that, ideally, *estimation procedures* should contribute relatively little to an index's variation.

Third, fit indices are designed to provide information about the degree to which a model is correctly or incorrectly specified for the given data. Obviously, *model misspecification* should directly affect fit indices. Put differently, the degree of model misspecification should be the major contributor to the variation of a sample fit index.

Finally, as in any other statistical estimation, two criteria apply in assessing the relative performance of competing estimators: *unbiasedness* and *variation*. Between two estimators, the one less biased is most often preferred; between two equally unbiased estimators, the one with less random variation is most often preferred. This consideration leads to two additional expectations: (a) a good fit index should have as little systematic bias (upward or downward) as possible; and (b) an ideal fit index should have as little random variation as possible.

Given the five criteria, relative performance of fit indices can be assessed through Monte Carlo experiments. Monte Carlo simulation becomes necessary mainly due to lack of theory with which to specify the distributions for the indices. As pointed by Bentler (1990), "Essentially nothing is known about the theoretical sampling distribution of the various estimators" (p. 245).

Previous Studies

Researchers have carried out simulation studies for most SEM model fit indices. Although some early studies focused on χ^2 behaviors under different sample size conditions (Boomsma, 1982), soon it became apparent that χ^2 test was too dependent on sample size to be useful in many situations. As a result, many alternative model fit indices were developed, and the majority of later simulation studies put more emphasis on these alternative model fit indices, especially those ranging from 0 to 1.

Invariably, all simulation studies investigated behaviors of model fit indices under different sample size conditions (Anderson

& Gerbing, 1984; Bearden, Sharma, & Teel, 1982; Bentler, 1990; Bollen, 1989; La Du & Tanaka, 1989; Marsh, Balla, & McDonald, 1988), since this has been considered a major weakness of the original χ^2 approach, and consequently, a major concern regarding alternative model fit indices. The majority of fit indices investigated, including the normed-fit-index (NFI), the goodness-of-fit index (GFI), and the adjusted goodness-of-fit index (AGFI), were shown to be influenced by sample size to different degrees.

But since different indices were involved in different studies, a performance comparison of the indices across different simulation designs tends to be difficult. Also, for obvious reasons, most studies looked at early fit indices, such as GFI, AGFI, NFI, etc., and some newer indices, such as McDonald centrality, Bollen's Delta2, etc., have rarely been investigated.

The sensitivity of some fit indices to model misspecification was examined in a few studies (Bentler, 1990; La Du & Tanaka, 1989; Marsh et al., 1988). The study by Marsh et al. (1988) was comprehensive in terms of the variety of fit indices studied, but the small number of replications in each cell condition (n=10) might have limited the generalizability of some conclusions. One finding from the study was that the relative fit indices, such as NFI, tended to be non-comparable across different studies or different data sets, since their values not only depended on model specification, but also, or more importantly, on how bad the null model itself was. Other studies (Bentler, 1990; La Du & Tanaka, 1989) involved fewer indices, making performance comparison among

fit indices difficult.

Very little is known about the influence of estimation methods on fit indices. In a few studies which examined the issue (La Du & Tanaka, 1989; Maiti & Mukherjee, 1991), maximum likelihood (ML) and generalized least squares (GLS) estimation procedures were used. Estimation procedures were shown to influence the value of the fit indices studied. But in these studies, only very few fit indices were examined, and the performance of other popular indices were unknown.

The purpose of the present study was to compare empirically the relative performance of SEM model fit indices. Three prominent factors which might affect SEM indices were considered: sample size, estimation procedure, and model misspecification. A three-factor experimental design was used to compare results across the Monte Carlo simulations. The variation of each fit index was partitioned to assess the influence of the three factors, and an index's behavior pattern was empirically examined.

METHOD

SEM Fit Indices Studied

Although a variety of fit indices exist, some of them are not readily comparable with each other. For example, Akaike's information criterion (AIC) both has such a different metric from many other fit indices and is used in such a different fashion that a meaningful comparison between AIC and GFI is difficult. Based on the consideration of comparability, eight popular indices were chosen for the study: goodness-of-fit index (GFI), adjusted

goodness-of-fit index (AGFI), Bentler's comparative fit index (CFI), McDonald's centrality index (Centrality), non-normed fit index (N_NFI), normed fit index (NFI), Bollen's normed fit index rho1 (RHO1), and Bollen's non-normed index delta2 (DELTA2). All these fit indices have an approximate range from zero to one, with higher values indicating better fit, and lower values indicating poorer fit. The comparable scale of these indices makes the comparison among them more straightforward.

Design of Monte Carlo Simulations

A three-factor balanced experimental design was used. The design is graphically represented in Figure 1. Five levels of the sample size condition (50, 100, 200, 500, 1000), three levels of model specification (true model, slightly misspecified model, moderately misspecified model), and two estimation methods (maximum likelihood, generalized least squares) were incorporated in the 5x3x2 design. Under this design, a total of 6,000 (5x3x2 x 200) replications of SEM model fitting were conducted, with 200 replications in each cell condition. Such a design allowed a systematic assessment of the impact of the three factors on fit indices: sample size, degree of model misspecification, estimation procedure. Also, 200 replications within each of the conditions provided estimates precisely enough to allow systematic comparisons among the fit indices on characteristics such as unbiasedness and degree of random variation.

Insert Figure 1 about here

Model and Model Specification

An SEM model of moderate complexity was simulated in the present study, as presented in Figure 2. This model was derived from a substantive research example described in the LISREL (Version 7) manual (Jöreskog & Sörbom, 1989, p. 178). As suggested elsewhere (Gerbing & Anderson, 1993), simulating substantively meaningful models in Monte Carlo simulation may increase the external validity of Monte Carlo research results.

Insert Figure 2 about here

The degree of SEM model complexity is a characteristic which is difficult to define, since complexity depends not only on the number of observed variables, but also on the number of latent variables, as well as on the unique relationship pattern among both given observed and latent variables. Most substantive studies using SEM involved from two to six latent variables, with about two to six indicators for each latent variable (Gerbing & Anderson, 1993). If this observation is correct, the model simulated in the present study, with four latent variables (two exogenous and two endogenous latent variables), each of which has three or four indicators, could be characterized as having moderate complexity, though of course such characterization is inherently subjective.

The population parameters for the true model in Figure 1 were artificially specified, as presented in Table 1 using LISREL representation. The population covariance matrix for the true model was obtained by using the prespecified parameters in Table

1 to reproduce the population covariance matrix, using the formula (Jöreskog & Sörbom, 1989, p. 5):

$$\Sigma = \begin{bmatrix} \Lambda_y (I-B)^{-1} (\Gamma\phi\Gamma'+\psi) (I-B')^{-1} \Lambda_y' + \theta_\epsilon & \Lambda_y (I-B)^{-1} \Gamma\phi\Lambda_x' \\ \Lambda_x\phi\Gamma' (I-B')^{-1} \Lambda_y' & \Lambda_x\phi\Lambda_x' + \theta_\delta \end{bmatrix}$$

Insert Table 1 about here

Table 2 presents the population covariance matrix reproduced using SEM population parameters in Table 1 and the formula above. Mathematically it is guaranteed that, other than rounding errors, perfect fit would be obtained if the model in Figure 2 was fit to this population covariance matrix. Since variable means do not affect SEM model fitting, to simplify the data generation process, all variables were centered with means being zeros. All sample data sets were generated based on the population covariance matrix in Table 2.

Insert Table 2 about here

Although a true model is relatively easy to specify in simulation research, model misspecification is difficult to handle for at least two reasons: (a) model misspecification can take such a variety of forms; and (b) the degree of model misspecification is not easily quantified, so it is difficult to make a priori prediction about the severity of misspecification (Gerbing & Anderson, 1993). We do not yet have solutions to these issues. In

the present study, model misspecification was achieved by fixing some parameters in the measurement model which should be set free, i.e., by setting some parameter values to be zero when, in fact, they were not, as indicated in Figure 2.

The degree of model misspecification was empirically determined by fitting two misspecified models to the population covariance matrix data, and the resultant values of fit indices were used as indicators of severity of model misfit. The terms "slightly misspecified" and "moderately misspecified" are used in the present paper simply to indicate different degrees of misspecification in this study; by no means should these terms be generalized beyond the present study, unless the degrees of misspecification are operationalized in the same manner.

Data Source

The present study only considered sample data generated from multivariate normal distributions. As a result, any issues related to data non-normality were not investigated. Data generation was accomplished using the data generator under the Statistical Analysis System (SAS PC Window Version 6.08). For each of the 6,000 sample data sets generated, the following steps were implemented:

(1) random normal variables with a desired sample size were generated, using the pseudorandom number generator under SAS;

(2) the random normal variables were linearly transformed to have desired means and standard deviations;

(3) the uncorrelated variables were then transformed to multivariate sample data with pre-specified population inter-variable correlations, using the matrix decomposition procedure (Kaiser & Dickman, 1962; Vale & Maurelli, 1983).

(4) The multivariate sample data was fit to one of the three models under one of the two estimation procedures, using PROC CALIS procedure under SAS. All desired fit indices from the sample were obtained and saved for later analysis.

Simulation programming was implemented through a combination of SAS Macro language, SAS PROC IML matrix language, and the SAS PROC CALIS procedure which implements structural equation modeling under the SAS environment. All simulation was carried out on an IBM PC Pentium 100 Mhz computer with SAS Windows Version 6.08.

Analysis

The major analytic strategy was to partition variation of sample fit indices into different components to assess the influence of different factors considered in the design. Since the design was a balanced experimental design, the variations due to different sources were orthogonal, which made the analysis and interpretation more straightforward. Factorial analysis of variance (ANOVA) was used as the analysis technique. This analysis allows us to partition the variation of a particular fit index into four major independent sources: sample size, estimation procedures, model misspecification, and random variation, plus some interaction

terms. Using the four criteria discussed previously, the behaviors of the eight fit indices were systematically examined, and their relative performance judged.

Besides partitioning sampling variance of the fit indices to assess the influences of different factors, values of fit indices were examined to assess characteristics such as the existence, or lack thereof, of systematic bias, and the extent of random variations for different indices.

RESULTS AND DISCUSSION

As discussed previously, five criteria were suggested for judging the relative performance of the eight fit indices examined in the present study: (a) sensitivity to sample size, (b) sensitivity to estimation methods, (c) sensitivity to model specification, and (d) degree of unbiasedness and (d) degree of random variation. Since there does not seem to be any consensus in the literature regarding the relative importance of these five features, the order of discussion of these issues should not be interpreted as reflecting implied relative importance of the criteria.

Table 3 presents descriptive data for the eight fit indices under different conditions: model specification, sample size conditions, estimation methods. Although more detailed data were available for the sample fit indices, e.g., confidence intervals, distribution characteristics (skewness, kurtosis), range, etc., here we present the basic information of means and standard deviations.

Insert Table 3 about here

Estimation Theory

An examination of Table 3 reveals several phenomena. First, under the true model (Model 1), the population values of the fit indices were essentially the same based on the two methods: maximum likelihood (ML) and generalized least squares (GLS). Also, under the true model, the sample means of the fit indices under different sample sizes (50, 100, 200, 500, 1000) were roughly comparable, especially with the increase of sample size. Three indices (CENTRA, NFI, RHO1) seemed to be exceptions. For these three indices, the two estimation methods exhibited noteworthy differences, especially under small sample sizes. For example, under a sample size 50, the mean for RHO1 under ML was .9006, while the mean under GLS was .9948. Similar differences occurred for CENTRA and NFI. With increased sample size, the difference of mean values between the two estimation methods seems to disappear. So *under the true model, both the population values and the sample means of the fit indices gave the impression that the two estimation methods in SEM provide comparable information about model fit, especially when sample size is reasonably large (e.g., over 200).*

However, under the two misspecified models (Model 1: slightly misspecified; Model 2: moderately misspecified), we observed some discrepancies between the two estimation methods both in terms of the population values of fit indices, and in terms of their sample

means under different sample size conditions. Under Model 2, such discrepancies did not appear to be too large, except for sample means of some individual fit indices (e.g., CENTRA, NFI, RHO1) under smaller sample size conditions (50, 100). Wherever such discrepancies occurred, fit index values based on GLS invariably *exceeded* those based on ML.

Under Model 3 (the moderately misspecified model), some discrepancies between the two estimation methods became alarmingly large, to the extent that they would give very different impressions about model fit. For example, for GFI, population values based on ML and GLS were .7902 and .9473, respectively; the population values for AGFI based on the two methods were .6791 and .9195, respectively. By current standards of model fit, the former values in both pairs would be judged as indicating very poor fit, while the latter values would be construed as indicating reasonable fit.

The same pattern occurred to varying degrees for the eight fit indices, and especially for the GFI, AGFI and CENTRA indices. Again, wherever discrepancies occurred, fit values based on GLS exceeded those based on ML, and in some cases, to considerable degrees. Such large discrepancies between the two estimation methods was not anticipated. Thus, *the two estimation methods seem to provide somewhat dissimilar information about model fit in the presence of model misspecification.*

Model Specification

Besides the comparison between the two different estimation

methods under different conditions, several other phenomena also stand out. One such phenomenon was the discrepancy in index performance across model specification conditions. Although different fit indices seemed to provide similar information about model fit under the true model, such was not the case for the two misspecified models. For example, for the slightly misspecified model, McDonald's centrality had a population value of .8714, while CFI and DELTA2 had values as high as .9798.

This situation became worse under the moderately misspecified model: GFI, AGFI, and CENTRA had population values of .7902, .6791, and .6087, respectively, while CFI, NFI, and DELTA2 had values of .9272, .9269, and .9272, respectively, under the same method (ML). Using conventional criteria for judging model fit, these two sets of fit indices would suggest very different conclusions about model fit, with the former group suggesting poor or very poor fit, and the latter group suggesting reasonably good fit. The difference across fit indices occurred in similar degrees for the sample means of fit indices under different sample size conditions, as well as under different estimation methods, i.e., under both ML and GLS.

These results suggest that *the fit indices were differentially sensitive to model misspecification.* As indicated by data in Table 3, GFI, AGFI, and CENTRA were more sensitive to model misspecification than the other five indices, all of which are relative fit indices, i.e., they are constructed by comparing the fit of the specified model with that of a null model.

Sampling Bias

Another observation based on data in Table 3 is related to systematic sampling bias of fit indices. It can be seen that *most sample fit indices tended to be systematically biased downward, though to different degrees*. For example, under the true model (Model 1), under sample sizes of 50, 100 and 200, GFI and AGFI showed fairly strong downward bias under both estimation methods, with sample means considerably lower than population values. The same was also true under the two misspecified models. Other fit indices exhibited similar downward bias pattern to lesser degrees. Of the eight fit indices, DELTA2, N_NFI, and CFI showed relatively slight downward sampling bias.

Sampling downward bias under the true model was expected, due to ceiling effect of fit indices. But the degree of downward bias of a few indices under misspecified models somewhat exceeded our expectations. Also, stronger downward bias seemed to occur for those indices which showed more sensitivity to model misspecification. More specifically, GFI, AGFI, and CENTRA showed more severe downward bias than the other fit indices under ML estimation. Furthermore, other than GFI and AGFI, downward bias seems to have only occurred under ML estimation method, but not under GLS estimation.

This absence of downward bias when using the GLS estimation method is probably related to the fact that GLS estimation tended to provide almost maximum fit index values even under Model 3 (moderately misspecified model). As a result, very little sampling variation occurred. A comparison of standard deviations between ML

and GLS estimation methods for the five indices (CFI, N_NFI, NFI, RHO1, and DELTA2) indicates substantially smaller standard deviations under GLS than those under ML, as reported in Table 3.

Sources of Variation in the Fit Statistics

Table 4 presents an ANOVA partitioning of the sampling variance of fit indices into different sources. The row labelled total sum of squares (SOS) provides an indication of the sampling variations of different indices. As indicated by the total sums of squares presented, the variation of fit indices differed substantially under the simulation conditions represented in the study: while the total SOS across all conditions for CFI was 5.118, the SOS for CENTRA was 131.9796, a difference of 30 times! *Three indices--GFI, AGFI, and CENTRA--which were shown earlier to be more sensitive to model misspecification than the other five indices, seem to have substantially larger variation than the other five.*

Insert Table 4 about here

Model Misspecification. The η^2's for model specification reported in Table 4, that is, the proportion of variance associated with model specification, indicates that CENTRA had the highest proportion in its variation (50.3%) which was contributed by model misspecification, while GFI (37.3%), AGFI (34.1%) followed, in that order. As reasoned before, since an fit index is designed to provide information about model fit, model specification (including model misspecification) should be a major contributor to an index's total variation. Also, large variation due to model specification

indicates an index's sensitivity to model misspecification. NFI, RHO1,and DELTA2 seemed to be least sensitive to model misspecification, as indicated by to small η^2s (15.1%, 16.3%, and 13.8%) for the condition of model specification. Based on this criterion, CENTRA would be ranked at the top, followed by GFI and AGFI. NFI, RHO1, and DELTA2 would be ranked at the bottom.

Sample Size. Sample size condition strongly influenced GFI and AGFI, accounting for 31.5% and 34.3% of total variance, respectively, for these two indices. CENTRA was the index least susceptible to sample size condition, with only .06% of variance accounted for by this condition. CFI and N_NFI also had very small percentages of total variance accounted for by sample size condition (.6% and .5% respectively). The other three fit indices had about 10% of total variance due to sample size. *The CENTRA index was least influenced by sample size, followed by CFI, N_NFI, while GFI and AGFI seemed to be overly affected by sample size.*

Estimation Method. Although the GFI and AGFI indices seemed to be overly influenced by sample size, *GFI and AGFI were least influenced by estimation method*, with about 10% of their total variation contributed by this factor. CENTRA followed GFI and AGFI in this regard, with about 20% of total variation accounted for by estimation method. The other five indices seemed to be overly affected by estimation method, with the percentage of total variation contributed by this factor ranging from 32.8% to 46.8%. Based on the criterion that a fit index should not be overly affected by estimation method, GFI and AGFI would be ranked best,

with CENTRA following these two. Again, NFI, RHO1, and DELTA2 would be worst.

Specification-by-Estimation Interaction. As reported in Table 4, the interaction term between model specification (MS) and estimation method (ES) accounted for a moderately large proportion of total variances for all the fit indices, ranging from 12.2% to 26.1%. This indicates that *model specification may have a stronger influence on fit indices under one estimation method than under another*.

A close look at Table 3 suggests that this interpretation is probably correct: model specification had much stronger influences on the estimated fit indices under ML than under GLS. As a matter of fact, for five indices (CFI, N_NFI, NFI, RHO1, and DELTA2), model misspecification seemed to have no impact at all on the estimated fit indices under GLS, with all these five indices attaining almost maximum values even under Model 3 (moderately misspecified model).

In other words, under GLS estimation, these fit indices are almost totally insensitive to the model misspecification conditions implemented in the present study; and their values gave the impression that even the moderately misspecified model was a model with perfect fit to the data. These findings were unexpected, and they raise serious questions about the effectiveness of these fit indices in providing information about SEM model fit, especially under GLS estimation.

Random Variation

Table 5 presents data on fit indices' random variation. Random variation was assessed through coefficients of variation (CV), which is considered a scale-free measure of variation. Using CV to represent random variation has the advantage of avoiding the problem of noncomparability across variables caused by different measurement metrics. CV is constructed as a ratio of sample standard deviation (s) to sample mean ($\bar{X}$) in percentage terms, with higher values representing more variation, regardless of measurement metric.

The results presented in Table 5 support several observations. First, *larger sample size resulted in smaller variation for all indices*. This was expected, since sampling variation should decrease with the increase of sample size. Second, *some indices tended to have considerably larger random variation than others*. Leading the list of fit indices in this regard was CENTRA, with consistently higher CVs than other indices under almost all conditions (models, sample sizes, estimation methods). Third, among the three models, *the severity of model misspecification resulted in larger random variation of fit indices*. Consistently, fit indices had larger random variation under the moderately misspecified model than that under slightly misspecified model, which, in turn, had larger variation than that under the true model.

Estimation method also seems to have a strong effect on fit indices' random variation. Invariably, *random variation was substantially larger under ML estimation than under GLS estimation*.

A few indices, e.g., CFI, N_NFI, NFI, DELTA2, had almost no random variation at all under GLS estimation. Although small random variation is generally considered as a positive aspect for a statistic, we suspect that the highly restricted random variation under GLS, especially for fit indices CFI, N_NFI, NFI, RHO1, and DELTA2, was caused by a ceiling effect of these fit indices estimated under GLS. If we look back at Table 3, it can be seen that these five indices under GLS estimation always attained almost maximum values, under the various sample size and/or what model specification conditions.

Conclusions

These results raise two important issues in SEM analysis. In most SEM applications, the major purpose is theory testing. This purpose is realized by examining how the predicted relationship pattern based on a theory can be supported by empirical data. In other words, the fit between a theoretical model and empirical data is of paramount importance in SEM analysis. Unfortunately, model fit as a central question in SEM analysis appears to be difficult to address, to say the least.

The first major issue raised by the results of the present study concerns the comparability of fit indices. The majority of previous Monte Carlo studies focused on correctly specified models, and much less empirical work has examined misspecified model of varying degrees. For a correctly specified model, fit indices seem to be comparable in that they all indicate that model fit is close to being perfect under reasonable sample size conditions. But for

misspecified models, the picture is different.

As indicated by the results of the present study, *fit indices may be much less comparable to each other than most researchers realize*. For example, using ML estimation, for our Model 3 (moderately misspecified model), under the sample size condition of 500 (reasonable sample size), a mean value of .7821 for GFI, as reported in Table 3, would certainly convey very different meaning about model fit from that based on a mean value of .9200 for NFI, or .9269 for DELTA2. Such discrepancies among fit indices have not been previously documented.

It is our belief that too much previous simulation research has focused on the true model, and not enough empirical work has been done for misspecified model. As a result, this comparability issue among fit indices has previously been largely ignored. Based on the results obtained in the present study, at least for the model conditions examined in the study, some fit indices appear to be much more sensitive to model misspecification (e.g., McDonald's Centrality, GFI, AGFI) than others (e.g., CFI, N_NFI, NFI, RHO1, DELTA2).

The second major issue involves intra-index comparability under different estimation methods. Theoretically, under multivariate normality conditions, ML and GLS estimation are asymptotically equivalent under large sample conditions (Gerbing & Anderson, 1993). If this is the case, empirically, we would expect that the discrepancy between fit indices' values under the two estimation methods would diminish as sample size increases. This

expectation, however, did not materialize. For example, for our Model 3 (moderately misspecified model), even under sample size condition of 1000, the mean value for GFI were .7850 and .9400, respectively, under ML and GLS estimation, as reported in Table 3. In research practice, such different fit index values based on the same model could lead to very different conclusions about model fit.

Similar intra-index discrepancy existed for other fit indices examined in the study. Here again, the discrepancy between estimation methods does not seem to be that obvious for models with less severe misspecification. Therefore, it is likely that this issue has been largely ignored in the literature due to the fact that previous focus has been on correctly specified models, and not enough work has been done for misspecified models.

We asked, which fit index has relatively better performance under different conditions? Although the results of the present empirical study does not provide the final answer to this question, some tentative conclusions can be presented.

To the extent that a fit index should be sensitive to model misspecification, the McDonald centrality index performed best, followed by GFI, and AGFI, with others trailing behind these three. If we desire an index which is minimally influenced by sample size, then the centrality index again came out to be the choice, followed by CFI, N_NFI, and some others. *As regards sensitivity to sample size, the GFI and AGFI--two very commonly used indices--did not perform well*, since sample size conditions accounted for more than

30% of their variations.

It is interesting to note that, in Tanaka's typology of fit indices (Tanaka, 1993), GFI, AGFI, and CFI were classified sample size dependent, while DELTA2 was classified sample size independent. Though this classification was empirically supported for GFI and AGFI, empirical results in the present study contradicted Tanaka's as regards both CFI and DELTA2: for CFI, only less than one tenth of a percent of variation was attributed to sample size conditions, while more than ten percent of variation was attributed to sample size conditions for DELTA2.

If we desire indices which are not overly influenced by estimation methods (ML or GLS in the present study), the GFI and AGFI indices seem to be the primary candidates, since their proportion of total variation which can be attributed to estimation method was appreciably less than other indices. This result, however, is based on estimation-appropriate GFI and AGFI indices, since different weight matrices are used in ML and GLS estimation to construct GFI and AGFI (SAS Institute, 1990; Tanaka, 1993). The centrality index trails GFI and AGFI in this regard. Other indices had considerably larger proportions of their variation associated with estimation method.

Downward bias occurred for almost all the fit indices examined, and such downward bias is more severe under smaller sample size conditions. For example, for our Model 1 (true model), under sample size 100, the 90% confidence interval (not presented in tables) for GFI under ML estimation would be (.9056, .9454),

while the population GFI was 1.0000. Although such downward bias is expected under the true model due to the ceiling effect of fit index values, similar downward bias also existed for misspecified models, as can be seen from Table 3. Other fit indices exhibited similar downward biases in varying degrees. The existence of such downward bias suggests that sample fit indices tend to present a somewhat more pessimistic picture about model fit than what is true in reality, especially when sample size is small. *Among the fit indices examined here, GFI and AGFI had the most serious downward bias under smaller sample sizes.*

Limitations

Several limitations of the present study should be acknowledged, since they may limit the generalizability of this study. The first limitation of the present study was that only one model was used as the basis for model specification condition, instead of a range of models varying in characteristics such as model complexity and different patterns of coefficient values in the model. Since only one model was used in the simulation, it is unknown to what extent the results can be replicated for other models, or for SEM analysis in general. The contributions of the present research must be augmented by further research.

The second limitation involves the precision of the estimates in the study. In the present study, after a sample was generated, the sample was fit to one model under one estimation method only, and samples in each cell were independent. For example, the 200 samples in the cell of sample size of 100, ML estimation, and True

Model specification were different from the other 200 samples in the cell of sample size of 100, ML estimation and slight model misspecification. The difference between these two cell conditions might be due to both model specification and sampling error. Although such confounding of model specification and sampling error may not be a statistical problem in the long run, it may affect the precision of study results if the number of samples in each cell is not large enough. To avoid such potential confounding, one sample could be generated and fit to all three models, instead of three independent samples being generated (Gerbing & Anderson, 1993).

SUMMARY

A balanced 3×2×5 (three model specification conditions × two estimation methods × five sample size conditions) design was used in a Monte Carlo simulation study to investigate the effects of these factors on SEM fit indices, with 200 replications within each cell. A total of 6,000 samples were generated from a prespecified population covariance matrix, and each of three prespecified models with known specification error were fit to data. Eight popular SEM fit indices were studied. The results of the present study suggest the following:

1. *Although fit indices seem to be comparable in providing information about model fit for correctly specified models, some fit indices appear to be non-comparable for incorrectly specified models.* Some fit indices seem to be much more sensitive to model misspecification than others, at least for the model conditions investigated in our study. This problem

has not been well documented in the literature, probably because previous studies focused more on correctly specified models, and not enough emphasis has previously been put on misspecified models.

2. *Estimation method (maximum likelihood versus generalized least squares in this study) strongly influence almost all the fit indices examined.* This influence does not seem to be obvious for correctly specified or slightly misspecified models; for more severe model misspecification, however, the effect appears to be strong. Again, this phenomenon has not previously been well documented in the literature. We suspect that the focus of previous studies on correctly specified model, rather than on misspecified model in SEM research, may have camouflaged this potential difference.
3. To recommend some fit indices at the expense of others is always difficult, since it is never certain if one particular study, or even a group of studies, has really captured the complexity of model fit within SEM analyses. Nevertheless, with this caveat in mind, based on the somewhat limited results of the study, *we tentatively recommend use of McDonald's Centrality, followed by GFI and AGFI indices, mainly for their sensitivity to model misspecification.* Other indices seem to have too little variance under different model specification conditions.

Obviously, more research is needed to address the important issues raised in the present study. We suggest that future

research should not only examine a wider range of models in terms of model complexity and some other characteristics, but also study a wider range of model specification conditions, including both true model and misspecified models of varying degrees.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Bearden, W. O., Sharma, S., & Teel, J. R. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. Journal of Marketing Research, 19, 425-530.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238-246.

Bentler, P. M. (1992). EQS structural equations program manual. Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1994). Forward. In B. M. Byrne, Structural Equation Modeling with EQS and EQS/Windows. Thousand Oaks, CA: SAGE Publications.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. Annual Review of Psychology, 47, 563-592.

Bollen, K. A. (1986). Sample size and Bentler and Bonnett's nonnormed fit index. Psychometrika, 51, 375-377

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. Sociological Methods & Research, 17, 303-316.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure,

prediction (Part I) (pp. 149-175). Amsterdam: North-Holland.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structure analysis and the "problem" of sample size: A clarification. Psychological Bulletin, 109, 512-519.

Gerbing, D. W. & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models. Newbury Park, CA: SAGE Publications, Inc.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis: Models, assumptions, and data. Beverly Hills, CA: Sage.

Jöreskog, K. G. & Sörbom, D. (1989). LISREL 7: A guide to the program and applications (2nd ed.). Chicago, IL: SPSS Inc.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 27, 179-182.

La Du, T. J., & Tanaka, J. S. (1989). The influence of sample size, estimation method, and model specification on goodness-of-fit assessment in structural equation models. Journal of Applied Psychology, 74, 625-635.

Maiti, S. S., & Mukherjee, B. N. (1991). Two new goodness-of-fit indices for covariance matrices with linear structure. British Journal of Mathematical and Statistical Psychology, 44, 153-180.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. Journal of Classification, 6, 97-103.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stillwell, C. D. (1989). An evaluation of goodness of fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

SAS Institute, Inc. (1990). SAS/STAT user's guide, version 6 (vol. 1, 4th ed.). Cary, NC: SAS Institute Inc.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), Sociological methodology 1985 (pp. 152-178). San Francisco, Jossey-Bass.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long, (Eds.), Testing structural equation models. Newbury Park, CA: SAGE Publications, Inc.

Tanaka, J. S., Huba, G. J. (1989). A fit index for covariance structure models under arbitrary GLS estimation. British Journal of Mathematical and Statistical Psychology, 42, 233-239.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B., & Daniel, L.G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. Educational and Psychological Measurement, 56,

213-224.

Vale, C.D. & Maurelli, V.A. (1983). Simulating multivariate nonnormal distributions. Psychometrika, 48, 465-471.

Table 1: Population Parameters for the True Model

$$\Lambda_x = \begin{bmatrix} .90 & .00 \\ .80 & .60 \\ .80 & .80 \\ .00 & 1.2 \\ .00 & 1.0 \\ 1.0 & .00 \end{bmatrix} \qquad \Lambda_y = \begin{bmatrix} .90 & .00 \\ .70 & .60 \\ .70 & .70 \\ .00 & .90 \\ .00 & 1.0 \\ 1.0 & .00 \end{bmatrix}$$

$$\Gamma = \begin{bmatrix} 1.10 & .00 \\ .00 & 1.00 \end{bmatrix} \qquad \Phi = \begin{bmatrix} 105.00 & 90.00 \\ 90.00 & 115.00 \end{bmatrix}$$

$$B = \begin{bmatrix} .00 & .00 \\ .00 & .00 \end{bmatrix} \qquad \Psi = \begin{bmatrix} 18.00 & .00 \\ .00 & 2.00 \end{bmatrix}$$

$$\theta_\delta = \begin{bmatrix} 30.00 \\ 30.00 \\ 40.00 \\ 45.00 \\ 20.00 \\ 50.00 \end{bmatrix} \qquad \theta_\varepsilon = \begin{bmatrix} 25.00 \\ 40.00 \\ 50.00 \\ 40.00 \\ 20.00 \\ 70.00 \end{bmatrix}$$

Table 2: Population Covariance Matrix for Generating Samples

σ^2	115.050	225.000	296.000	210.600	135.000	155.000	142.491	236.354	275.425	134.770	137.000	215.051
μ	0	0	0	0	0	0	0	0	0	0	0	0
X1	1.0000											
X2	.771946	1.00000										
X3	.760813	.864882	1.00000									
X4	.624444	.777284	.788224	1.00000								
X5	.649942	.809023	.820410	.818432	1.00000							
X6	.707656	.738961	.728304	.597762	.622171	1.00000						
Y1	.730685	.763009	.752005	.617215	.642418	.699464	1.00000					
Y2	.735983	.827640	.826158	.743861	.774236	.704536	.789256	1.00000				
Y3	.727290	.823335	.822759	.746384	.776861	.696214	.776113	.824262	1.00000			
Y4	.585447	.728741	.738999	.737217	.767320	.560431	.578669	.703458	.706311	1.00000		
Y5	.645180	.803096	.814399	.812436	.845611	.617612	.637711	.775232	.778377	.774946	1.00000	
Y6	.660863	.690098	.680145	.558235	.581030	.632625	.745757	.713836	.701949	.523373	.5767734	1.0000

Table 3: Means and Standard Deviations of Eight Fit Indices Under Two Estimation Methods

Fit Indices	GFI[1]	AGFI	CFI	CENTRA	N_NFI	NFI	RHO1	DELTA2
Model 1[2]	1.0000[3]	1.0000	1.0000	1.0024	1.0005	1.0000	1.0000	1.0003
Parameters	1.0000	1.0000	1.0000	1.0024	1.0001	1.0000	1.0000	1.0001
Sample Sizes								
50	.8629 (.0243)[4]	.7725 (.0403)	.9893 (.0122)	.9576 (.1058)	.9900 (.0232)	.9292 (.0158)	.9006 (.0227)	.9931 (.0161)
	.8508 (.0289)	.7524 (.0479)	.9998 (.0003)	1.0356 (.0870)	1.001 (.0012)	.9963 (.0015)	.9948 (.0021)	1.000 (.0009)
100	.9258 (.0118)	.8769 (.0196)	.9960 (.0048)	.9853 (.0445)	.9968 (.0094)	.9647 (.0069)	.9505 (.0096)	.9977 (.0066)
	.9255 (.0153)	.8764 (.0254)	.9998 (.0003)	1.015 (.0457)	1.000 (.0001)	.9969 (.0001)	.9957 (.0013)	1.000 (.0007)
200	.9627 (.0077)	.9382 (.0129)	.9985 (.0025)	.9993 (.0258)	.9998 (.0053)	.9831 (.0039)	.9762 (.0054)	.9998 (.0038)
	.9622 (.0069)	.9372 (.0115)	.9999 (.0002)	1.005 (.0208)	1.000 (.0001)	.9979 (.0005)	.9971 (.0007)	1.000 (.0004)
500	.9847 (.0030)	.9745 (.0050)	.9994 (.0009)	.9999 (.0096)	.9999 (.0020)	.9931 (.0014)	.9903 (.0020)	.9999 (.0014)
	.9849 (.0028)	.9750 (.0047)	.9995 (.0001)	1.002 (.0084)	1.000 (.0002)	.9991 (.0002)	.9987 (.0002)	1.000 (.0002)
1000	.9921 (.0016	.9870 (.0026)	.9997 (.0004)	.9999 (.0048)	.9999 (.0001)	.9965 (.0007)	.9951 (.0009)	.9999 (.0007)
	.9923 (.0015	.9873 (.0025)	.9999 (.0001)	1.001 (.0046)	1.000 (.0001)	.9995 (.0001)	.9994 (.0001)	1.000 (.0001)
Model 2[5]	.9475	.9165	.9798	.8714	.9728	.9795	.9724	.9798
Parameters	.9722	.9557	.9983	.9221	.9977	.9982	.9976	.9983
Sample Sizes								
50	.8142 (.0338)	.7043 (.0538)	.9646 (.0206)	.7984 (.1159)	.9528 (.0286)	.9009 (.0205)	.8665 (.0276)	.9659 (.0207)
	.8308 (.0286)	.7306 (.0455)	.9997 (.0004)	.9961 (.0833)	1.000 (.0012)	.9957 (.0020)	.9942 (.0027)	1.000 (.0009)
100	.8814 (.0242)	.8112 (.0385)	.9773 (.0124)	.8614 (.0723)	.9695 (.0168)	.9435 (.0121)	.9239 (.0162)	.9777 (.0123)
	.8986 (.0159)	.8386 (.0253)	.9992 (.0006)	.9465 (.0445)	.9990 (.0009)	.9958 (.0011)	.9943 (.0015)	.9993 (.0006)
200	.9139 (.0157)	.8629 (.0251)	.9791 (.0074)	.8691 (.0427)	.9718 (.0099)	.9618 (.0073)	.9486 (.0098)	.9792 (.0073)
	.9361 (.0101)	.8983 (.0161)	.9988 (.0006)	.9344 (.0281)	.9984 (.0007)	.9966 (.0001)	.9954 (.0011)	.9988 (.0006)
500	.9344 (.0086)	.8956 (.0137)	.9796 (.0035)	.8714 (.0205)	.9727 (.0047)	.9728 (.0035)	.9634 (.0047)	.9798 (.0035)
	.9578 (.0048)	.9329 (.0077)	.9985 (.0003)	.9258 (.0134)	.9979 (.0004)	.9975 (.0004)	.9966 (.0005)	.9984 (.0003)
1000	.9397 (.0072)	.9040 (.0115)	.9792 (.0028)	.8674 (.0162)	.9719 (.0038)	.9757 (.0028)	.9673 (.0038)	.9792 (.0028)
	.9647 (.0032)	.9438 (.0050)	.9983 (.0002)	.9220 (.0087)	.9977 (.0003)	.9978 (.0003)	.9970 (.0004)	.9983 (.0002)

(To be Continued)

Table 3: Means and Standard Deviations of Eight Fit Indices Under Two Estimation Methods (Continued)

Fit Indices	GFI	AGFI	CFI	CENTRA	N_NFI	NFI	RHO1	DELTA2
Model 3[4]	.7902	.6791	.9272	.6087	.9058	.9269	.9053	.9272
Parameters	.9473	.9195	.9967	.8561	.9957	.9966	.9956	.9967
Sample Sizes								
50	.7044 (.0540)	.5480 (.0826)	.9134 (.0337)	.5763 (.1197)	.8879 (.0435)	.8510 (.0329)	.8072 (.0427)	.9153 (.0328)
	.8039 (.0291)	.7001 (.0445)	.9994 (.0006)	.9390 (.0799)	.9994 (.0009)	.9949 (.0019)	.9934 (.0025)	.9995 (.0007)
100	.7503 (.0445)	.6181 (.0680)	.9257 (.0204)	.6109 (.0839)	.9039 (.0264)	.8931 (.0199)	.8617 (.0258)	.9266 (.0201)
	.8767 (.0169)	.8114 (.0259)	.9985 (.0007)	.8959 (.0445)	.9980 (.0009)	.9948 (.0014)	.9933 (.0018)	.9985 (.0007)
200	.7692 (.0337)	.6470 (.0515)	.9264 (.0147)	.6102 (.0574)	.9047 (.0189)	.9098 (.0146)	.8832 (.0189)	.9268 (.0146)
	.9126 (.0093)	.8663 (.0142)	.9976 (.0007)	.8754 (.0242)	.9969 (.0009)	.9954 (.0015)	.9940 (.0015)	.9976 (.0007)
500	.7821 (.0222)	.6667 (.0339)	.9268 (.0092)	.6097 (.0359)	.9052 (.0119)	.9200 (.0093)	.8965 (.0119)	.9269 (.0092)
	.9330 (.0050)	.8975 (.0076)	.9969 (.0005)	.8611 (.0129)	.9961 (.0007)	.9959 (.0006)	.9947 (.0008)	.9970 (.0005)
1000	.7850 (.0158)	.6711 (.0241)	.9267 (.0062)	.6067 (.0250)	.9051 (.0079)	.9234 (.0062)	.9008 (.0079)	.9268 (.0062)
	.9400 (.0029)	.9083 (.0043)	.9969 (.0003)	.8571 (.0073)	.9959 (.0004)	.9963 (.0003)	.9952 (.0004)	.9968 (.0003)

1 GFI: goodness-of-fit index; AGFI: adjusted goodness-of-fit index; CFI: comparative fit index; CENTRA: McDonald's centrality index; N-NFI: non-normed fit index; NFI: normed fit index; RHO1: Bollen's RHO1; DELTA2: Bollen's delta2 (incremental fit) index.

2 Model 1: correctly specified model (true model)

3 Population fit index value; the upper is based on maximum likelihood method, and the lower is based on generalized least squares method. These population values were obtained by fitting the relevant prespecified model to the population covariance matrix with N set at 10,000. For the two misspecified models, these populations values are in fact estimates. Since model fitting was based on population data, and the N was set to be sufficiently large, these values could be treated as proxies of population parameters.

4 Mean and standard deviation (SDs in parentheses) of a fit index; the values in the first row were based on maximum likelihood estimation, and those in the second row were based on generalized least squares estimation.

5 Model 2: slightly misspecified model

6 Model 3: moderately misspecified model

Table 4: Eta-Squares due to Different Sources for the Eight Fit Indices

Source	Fit Indices							
	GFI	AGFI	CFI	CENTRA	N-NFI	NFI	RHO1	DELTA2
Total SOS[1]	39.5208	92.9333	5.1118	131.9796	8.7602	10.1052	5.9067	17.3151
Model Specification	.3730[2]	.3413	.2902	.5030	.2918	.1512	.1628	.1381
Sample Size	.3154	.3426	.0059	.0006	.0049	.1079	.1018	.1143
Estimation Method	.0993	.0996	.3281	.1996	.3199	.4566	.4458	.4678
MS*SS	.0028	.0053	.0003	.0011	.0004	.0002	.0004	.0002
MS*EM	.1371	.1360	.2610	.1374	.2597	.1331	.1438	.1222
SS*EM	.0033	.0034	.0084	.0198	.0078	.0957	.0905	.1013
MS*SS*EM	.0024	.0023	.0004	.0013	.0007	.0003	.0006	.0002

1 Total sum of squares

2 η^2: proportion of variance contributed by a source; obtained through:

(sum of squares due to a source)/(total sum of squares)

Table 5: Random Variation of Fit Indices: Coefficients of Variation[1]

	Fit Indices															
	GFI		AGFI		CFI		CENTRA		N_NFI		NFI		RHO1		DELTA2	
	ML[2]	GLS	ML	GLS	ML	GLS	ML	GLS	ML	GLS	ML	GLS	ML	GLS	ML	GLS
50	2.82[3]	3.39	5.22	6.38	1.23	.03	11.05	8.40	2.34	.12	1.71	.15	2.47	.21	1.62	.09
	4.15	3.43	7.63	6.22	2.13	.04	14.53	8.35	3.01	.12	2.28	.20	3.19	.27	2.15	.09
	7.67	3.62	15.08	6.36	3.68	.06	20.77	8.50	4.91	.09	3.87	.20	5.28	.25	3.58	.07
100	1.28	1.65	2.23	2.89	.48	.03	4.52	4.49	.94	.09	.71	.09	1.01	.13	.66	.07
	2.75	1.76	4.74	3.01	1.26	.06	8.39	4.70	1.73	.09	1.28	.11	1.76	.15	1.26	.06
	5.93	1.92	11.01	3.19	2.20	.07	13.73	4.97	2.92	.09	2.23	.14	2.99	.18	2.17	.07
200	.80	.72	1.37	1.23	.25	.02	2.59	2.07	.54	.05	.39	.05	.56	.08	.38	.04
	1.72	1.08	2.91	1.79	.75	.06	4.92	3.01	1.02	.08	.76	.09	1.04	.11	.75	.06
	4.38	1.01	7.96	1.64	1.58	.07	9.40	2.77	2.10	.09	1.61	.12	2.14	.15	1.57	.07
500	.31	.29	.52	.48	.09	.01	.96	.84	.20	.02	.15	.02	.21	.02	.14	.02
	.93	.51	1.53	.83	.35	.03	2.36	1.45	.48	.04	.36	.04	.49	.05	.36	.03
	2.90	.54	5.08	.85	.99	.05	5.89	1.49	1.32	.07	1.01	.06	1.34	.08	.99	.05
1000	.16	.15	.26	.26	.05	.01	.48	.45	.10	.01	.07	.01	.10	.01	.07	.01
	.77	.33	1.27	.53	.28	.02	1.87	.95	.39	.03	.29	.03	.39	.04	.28	.02
	2.01	.30	3.59	.48	.67	.03	4.12	.86	.88	.04	.67	.03	.89	.04	.66	.03

1 Coefficients of Variation: CV = (standard deviation/mean) × 100

2 ML: maximum likelihood estimation; GLS: generalized least squares estimation.

3 The three coefficient of variation (CV) in the column are for the three models respectively: true model, slightly misspecified model, and moderately misspecified model, in that order.

Table ?: Influences of Model Specification and Sample Size under Two Estimation Methods (η^2)

	Fit Indices							
	GFI	AGFI	CFI	CENTRA	N-NFI	NFI	RHO1	DELTA2
ML								
Total SOS	26.0708[1]	60.2106	3.4309	90.0887	5.9493	5.4819	9.1995	3.4733
Model Specification	.7254[2]	.6971	.8206	.8502	.8112	.5234	.4893	.8241
Sample Size (SS)	.1930	.2147	.0211	.0119	.0183	.3748	.4054	.0149
MS*SS	.0078	.0110	.0008	.0016	.0015	.0009	.0007	.0013
GLS								
Total SOS	9.5262	23.4678	.0039	15.5496	.0086	.0091	.0155	.0050
Model Specification	.1308	.1014	.5391	.5100	.5128	.3391	.2892	.5211
Sample Size	.7938	.8192	.1763	.1040	.2114	.2600	.2836	.2088
MS*SS	.0005	.0023	.1327	.0110	.0697	.0263	.0343	.0737

1 Total Sum of Squares

2 η^2, obtained through: (sum of squares due to a source)/(total sum of squares)

Figure 1: Three-Factor Balanced Design

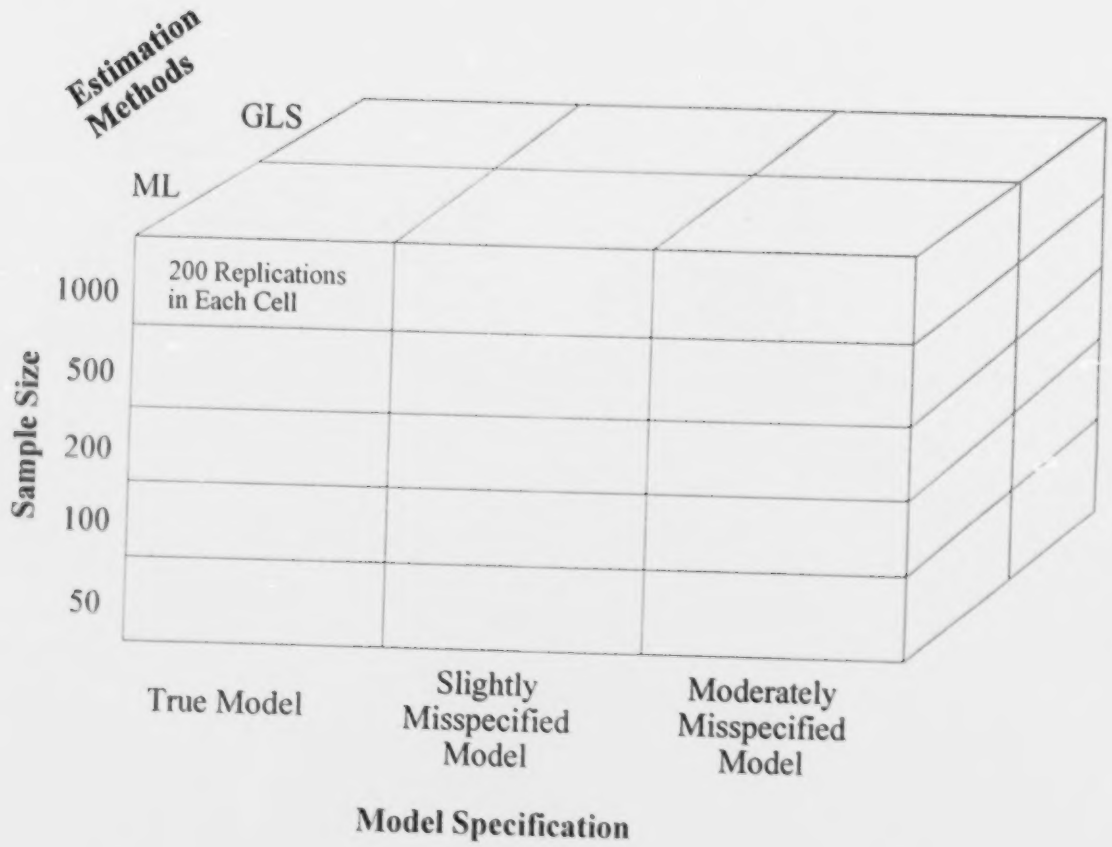
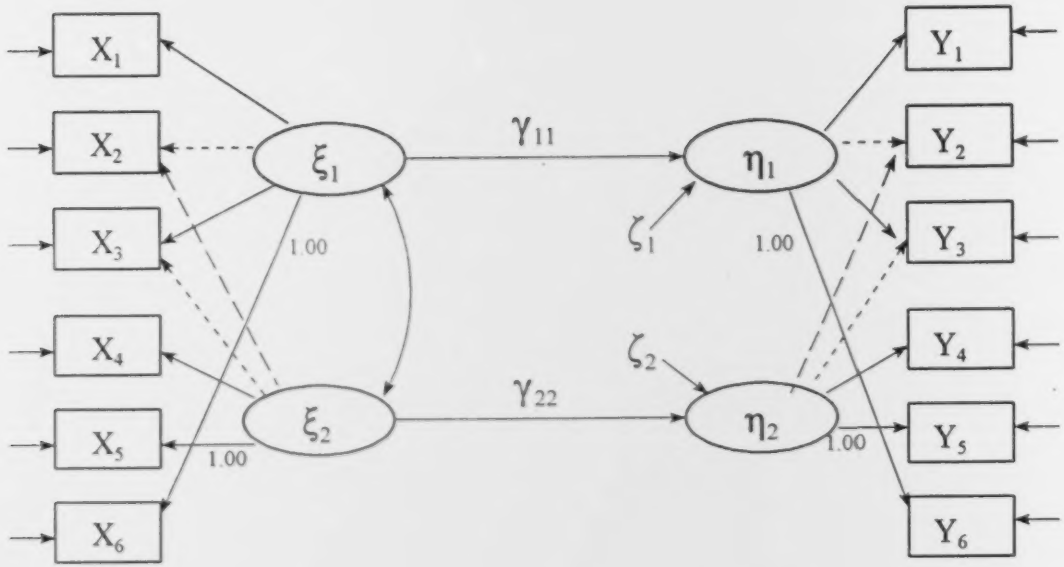

Estimation Methods
GLS
ML
200 Replications in Each Cell
1000
500
200
100
50
Sample Size
True Model
Slightly Misspecified Model
Moderately Misspecified Model
Model Specification

Figure 2: Simulated SEM Model: True Model and Two Misspecified Models



Paths fixed to be zero for slightly misspecified model.

Paths fixed to be zero for moderately misspecified model.